How speaker tongue and name source language affect the automatic recognition of spoken names

Bert Réveil[1], *Jean-Pierre Martens*[1], *Bart D'hoore*[2]

[1]DSSP, ELIS, Ghent University, Sint-Pietersnieuwstraat 41, 9000 Ghent, Belgium
[2]Nuance, Guldensporenpark 32, 9820 Merelbeke, Belgium

bert.reveil@elis.ugent.be, jean-pierre.martens@elis.ugent.be, bart.dhoore@nuance.com

Abstract

In this paper the automatic recognition of person names and geographical names uttered by native and non-native speakers is examined in an experimental set-up. The major aim was to raise our understanding of how well and under which circumstances previously proposed methods of multilingual pronunciation modeling and multilingual acoustic modeling contribute to a better name recognition in a cross-lingual context. To come to a meaningful interpretation of results we have categorized each language according to the amount of exposure a native speaker is expected to have had to this language. After having interpreted our results we have also tried to find an answer to the question of how much further improvement one might be able to attain with a more advanced pronunciation modeling technique which we plan to develop.

Index Terms: speech recognition, proper names, pronunciation modeling

1. Introduction

The automatic recognition of proper names in a car navigation or a directory assistance application is still a challenge. First of all, many involved names may exhibit archaic spellings or partly originate from foreign languages. Furthermore, non-native users often have to be accommodated. Consequently, transcriptions emerging from a native grapheme-to-phoneme (g2p) converter cannot capture the large variety of pronunciations that need to be dealt with [1].

A first attempt to resolve these difficulties [2, 3] consists of including transcriptions that were computed by foreign g2p converters. To comply with the monolingual acoustic models, the g2p outputs are *nativized* by mapping each foreign phoneme to its closest native equivalent. In [2], English and French transcriptions were included in a pronunciation dictionary containing Dutch, English, French and other names. The name error rate (NER) dropped by about 40% for native Dutch speakers, 70% for French speakers, 45% for English speakers and over 10% for the other foreign speakers. However, the vocabulary size was small ($<$ 500 entries) and therefore not representative of a car navigation task for instance. In [3], eight g2p converters (Mandarin Chinese, Czech, French, German, Hindi, Italian, Russian and Spanish) were available for creating extra foreign transcriptions for 24K proper names of various origins. A 25% reduction of the NER for foreign names spoken by speakers of the name language was achieved, as well as a 10% reduction for foreign names spoken by American English native speakers.

Another approach [4, 5, 6] is to create accented pronunciations by means of phonological rules. In [4], these rules were learned from alignments of the native (German) transcription with the outputs of a non-native (English) phoneme recognizer. In [5] and [6], manually compiled phonological rules were employed. The drops in NER were mostly moderate (5 to 15%) with some rare degradations as well.

Many authors [5, 7, 8] argued that including better phonetic transcriptions in the lexicon is not enough. One also needs acoustic models that can represent accented sounds that rarely occur in regular native training material.

In [5], non-native speech with nativized transcriptions was included in the training material of the acoustic models. In combination with the previously mentioned phonological rules the recognition of French words and expressions spoken by American English, German and Spanish speakers greatly improved: a 25% reduction of the error rate. There was also a reduction of 12.5% for the non-native speakers with other mother tongues.

Other authors [7, 8] worked with multilingual acoustic models trained on multilingual speech with multilingual transcriptions. In [8], the native language is German, and the foreign training material consists of English speech spoken by Germans. In a German spoken dialogue system that answers questions about (English) movies, theatres, timetables, etc., the NER dropped by 25% over that of a monolingual system with a knowledge-based mapping of English to German (native) phonemes.

In this paper we mainly investigate the most successful approaches as a function of name source (language of origin of the name) and speaker tongue (mother tongue of the speaker). We produce some new results which have raised our understanding of the mechanisms that are responsible for producing the observed recognition improvements. We also wonder how much our best known system could be further improved by means of more advanced pronunciation modeling.

2. Experimental set-up

For our experimental study we needed a spoken name database, g2p converters for several languages and a speech recognizer.

2.1. Spoken name database

Since we aim to investigate name source and speaker tongue as dependent variables, we opted for a corpus that is balanced with respect to these variables. The Autonomata Spoken Name Corpus (ASNC) [9] contains utterances of Dutch, English, French, Moroccan and Turkish person names (first name + family name) and geographical names (street names and city names) spoken by 120 Dutch, 40 English, 20 French, 40 Moroccan and 20 Turkish speakers. Speakers were recorded in two regions: Flanders and the Netherlands. Each speaker read 181 names: (1) 120 Dutch names (40 person names and 80 geographical names), (2)

23 English names (7 person names and 16 geographical names), (3) 15 Moroccan person names and (4) either 23 French names (in Flanders) or 23 Turkish names (in the Netherlands) (7 person names and 16 geographical names). The reason for the latter distinction is that many speakers in Flanders are familiar with French (the second language in Belgium) whereas speakers in the Netherlands are not.

There were 10 mutually exclusive name lists per region: 12 were read by 16 speakers, the other 8 by only 6. Because of a few overlaps between the lists that were used in Flanders and in the Netherlands, there were only 3540 different names, rather than the maximally expected 20 x 181 = 3620.

For experimentation, the corpus was divided in a train set and a test set, and there was no overlap in speakers nor name lists between both parts.

2.2. Phonetic transcriptions

For generating phonetic transcriptions we utilized the Dutch, English, French and German g2p converters that are embedded in the Nuance RealSpeak text-to-speech system[1]. The German g2p converter was also included because many Dutch speakers are familiar with German, and because we will test a multilingual recognizer which has seen German speech during training as well.

The ASNC comes with auditorily verified transcriptions (1 per name utterance). The auditorily verified transcription of an utterance is the best nativized transliteration of what a human expert actually heard after listening.

2.3. Recognition system

The experiments were executed with the commercially available Nuance VoCon 3200 recognizer[1]. In order to investigate the effect of moving from monolingual to multilingual acoustic models, the engine came with two sets of acoustic models:

- AC-MONO: the standard Dutch model, trained on speech of native Dutch speakers from the Netherlands and Belgium. The underlying phoneme set consists of 45 phonemes, and the size of the model is 2.7 MB.
- AC-MULTI: a multilingual acoustic model, trained on the same data as AC-MONO, supplemented with UK English, French and German speech. The Dutch portion now constitutes only 20% of the total training data. The underlying phoneme set consists of 80 phonemes. The size of the model is 4.5 MB and it contains roughly 70% more parameters than AC-MONO. Models for phonemes appearing in multiple languages have seen data from all these languages.

As a grammar we considered a loop comprising all 3540 names. As a performance measure we adopted the name error rate (NER), meaning that a name is only correct if all its constituents (words) are correct.

3. Experimental study

We argue that cross-lingual results can be explained efficiently by making a distinction between the target language of the application (Dutch in our case), hereafter called the native language (NAT), and two types of non-native languages. The first type (NN1) consists of languages (French and English in our case) whose pronunciation rules are known to many native speakers. The second type (NN2) consists of all other languages (Turkish and Moroccan in our case). This language distinction is used to discern the speaker tongue as well as the name source. Table 1 shows the emerging division of the test set.

Table 1: *Number of name utterances in the test set for the different name sources and speaker tongues*

	Name source			
speaker tongue	**NAT**	**NN1**	**NN2**	**All**
NAT	4400	1265	992	6697
NN1	2520	805	476	3801
NN2	2280	575	584	3439
All	9200	2645	2052	13937

3.1. Baseline system

Our baseline system uses monolingual acoustic models (AC-MONO) and Dutch g2p transcriptions. The NERs of this system are listed in Table 2. As expected, the recognition of native names by native speakers is already quite reliable but as soon as cross-lingual effects come into play, the NERs are substantially higher. Note that for all speaker categories the recognition of NN2 names is substantially better than that of NN1 names. We tested two hypotheses in this respect.

Table 2: *NER (%) obtained with AC-MONO and a lexicon with Dutch g2p transcriptions*

	Name source			
Speaker tongue	**NAT**	**NN1**	**NN2**	**All**
NAT	3.9	22.5	12.6	8.7
NN1	18.1	37.5	14.7	21.8
NN2	22.5	36.4	29.3	26.0
All	12.4	30.1	17.8	16.6

Our first hypothesis that speakers use Dutch g2p-knowledge to read these unfamiliar names was not supported by the data. The mean discrepancy (in phonetic symbol difference rate) between Dutch g2p transcriptions and auditorily verified transcriptions appeared to be the same for NN1 and NN2 names.

A second hypothesis was that NN2 names are easier to recognize because they bear less affiliation with the Dutch language. This hypothesis is confirmed by the fact that 60% of the misrecognized NN1 names are confused with a Dutch name, while for NN2 names only 40% are.

In the subsequent experiments we will assess how the NERs are affected by changes in the system configuration.

3.2. Experiment 1: adding nativized foreign transcriptions

In a first experiment, the Dutch g2p transcription of a name was supplemented with nativized French and English (NN1 languages) g2p transcriptions. The utilized phoneme mappings were suggested by a human expert who did not see the name lists.

Table 3 shows that the inclusion of foreign transcriptions helps a lot to improve the recognition of NN1 names (true for both English and French names). Somewhat surprisingly, we found that the gains for NN1 names uttered by NAT speakers were larger than for NN1 names uttered by native speakers of

[1] http://www.nuance.com/

the NN1 language in question. This differs from [2, 3] where the opposite result was obtained.

In about 95% of the cases where an improvement for English (French) names was found, the English (French) transcription was chosen. Our results thus support the idea advocated in [10] that Dutch speakers use their NN1 language knowledge for reading NN1 names. In that respect we see that most NN2 speakers also seem to use their knowledge of English (a world language) or French (second official language of Morocco).

Table 3: *NER (%) obtained with AC-MONO and a lexicon with Dutch + nativized English and French g2p transcriptions. In bold are gains > 20% w.r.t. the baseline system.*

	Name source			
Speaker tongue	**NAT**	**NN1**	**NN2**	**All**
NAT	4.0	**8.5**	11.3	**5.9**
NN1	16.6	**21.1**	**10.9**	**16.8**
NN2	22.9	**28.4**	28.1	24.7
All	12.1	**16.7**	16.0	13.5

The substantial NER gain for NN1 speakers reading NN2 names is owed to the fact that Moroccan names had a French spelling, thus calling for a French-like pronunciation. Finally, it is noted that the presence of foreign transcriptions does not hurt the recognition of native names by native speakers.

3.3. Experiment 2: multilingual acoustic models

Moving to multilingual acoustic models in combination with plain NN1 transcriptions (not nativized) leads to substantial additional improvements for NN1 speakers (see Table 4). As argued in [10], NN1 speakers can have a strong accent and produce sounds from their mother tongue. The recognition of these speakers thus improves by the availability of appropriate acoustic models for these sounds. This is most strong for the NN1 names which incorporate more of these sounds. The improvement rates found for English names read by English natives and French names read by French natives were 82% and 93% respectively.

Table 4: *NER (%) obtained with AC-MULTI and a lexicon with Dutch, English and French g2p transcriptions. In bold/italic are gains/losses > 20% w.r.t. the system of experiment 1.*

	Name categories			
Speaker categories	**NAT**	**NN1**	**NN2**	**All**
NAT	*4.9*	**5.9**	**8.9**	5.7
NN1	**11.6**	**7.1**	**6.9**	**10.0**
NN2	21.6	**21.0**	**20.0**	21.3
All	10.8	**9.6**	**11.6**	**10.7**

Two other results are that NN2 as well as NN1 names benefit from the multilingual acoustic model set (as in [5]) and that the NER for native speakers reading native names increases a lot. We owe the latter to the fact that many phonemes appear in different languages, and consequently, their acoustic models are 'contaminated' by foreign pronunciations.

In order to explain the first result we conducted a control experiment in which the multilingual transcriptions were replaced by the nativized transcriptions of experiment 1. In that case the recognizer cannot call its foreign phoneme acoustic models anymore, and this explains why the NER for the (NN1,NN1) combination increased from 7.1 to 9.8%. However, more surprisingly, there were no significant degradations for the other combinations. Apparently, the improvement for NN1 and NN2 names across speaker tongues mainly stems from the fact that multilingual acoustic models of native phonemes are less specialized in the recognition of native sounds, and therefore beneficial to the recognition of accented sounds.

We argue that including NN2 language data in the multilingual acoustic model training might further improve the recognition of NN2 names spoken by NN2 speakers, but we did not verify this yet.

3.4. Experiment 3: adding German transcriptions

Since the acoustic training data also comprised German speech, and since German can be considered as a NN1 language, we have also investigated the effect of adding a German g2p transcription per name to the lexicon of experiment 2. Table 5 shows that the recognition of native names spoken by native speakers further degrades, but that of NN2 names uttered by native speakers improves. An analysis of the previously incorrect but now correctly recognized NN2 names learned that most of them contain one or more occurrences of the character 'u' (Curukluk Sokagi, Butrus Benhida, Oglumus Rasuli,...). A Dutch g2p converter converts this character to /ˆ/ (like in m**u**d) or /y/ (no English equivalent, like in the French écr**u**) whereas a German g2p converter will more likely return a /u:/ (like in b**oo**t) or /U/ (like in b**oo**k). Apparently, many Dutch speakers pronounce these names with one of the latter phonemes.

Table 5: *NER (%) obtained with AC-MULTI and a lexicon with Dutch, English, French and German g2p transcriptions. In bold are gains > 20% w.r.t. the system of experiment 2.*

	Name source			
Speaker tongue	**NAT**	**NN1**	**NN2**	**All**
NAT	5.2	5.8	**6.0**	5.4
NN1	10.5	7.1	5.9	9.2
NN2	20.6	21.2	17.3	20.3
All	10.4	9.5	**9.2**	10.1

3.5. Experiment 4: using the name origin

If the name source were known, one could be more selective in adding foreign transcriptions, in the hope to eliminate the loss of accuracy for native names spoken by native speakers. Therefore we set up a test with Dutch transcriptions for Dutch names, Dutch and English transcriptions for English names, Dutch and French transcriptions for French names and all four transcriptions for NN2 names. Table 6 confirms our expectation, but it also shows that the recognition of native names by NN1 speakers further degrades, but that of NN2 names uttered by native speakers improves. An analysis of the previously incorrect but now correctly recognized NN2 names learned that most of them contain one or more occurrences of the character 'u' (Curukluk Sokagi, Butrus Benhida, Oglumus Rasuli,...). A Dutch g2p converter converts this character to /ˆ/ (like in m**u**d) or /y/ (no English equivalent, like in the French écr**u**) whereas a German g2p converter will more likely return a /u:/ (like in b**oo**t) or /U/ (like in b**oo**k). Apparently, many Dutch speakers pronounce these names with one of the latter phonemes.

Table 6: *NER (%) obtained with AC-MULTI and a lexicon with name source specific transcriptions. In bold/italic are gains/losses > 20% w.r.t. the system of experiment 3.*

	Name source			
Speaker tongue	**NAT**	**NN1**	**NN2**	**All**
NAT	**4.2**	5.3	5.3	4.9
NN1	*13.7*	7.6	5.0	*11.3*
NN2	21.5	20.4	14.2	20.1
All	11.3	9.3	7.8	10.4

kers substantially degrades. These speakers use their English / French knowledge to pronounce Dutch names.

3.6. Experiment 5: using the speaker origin

We also performed an experiment in which the transcription selection was guided by the speaker tongue: Dutch transcriptions for Dutch speakers, Dutch and English transcriptions for English speakers, Dutch and French transcriptions for French speakers and all four transcriptions for Morrocan and Turkish speakers. As anticipated, the recognition of foreign names spoken by native speakers went back to the level of the baseline system. The other combinations were less or not affected.

4. Pronunciation modeling

The former experimental study revealed what can be achieved with a pronunciation model based on existing general-purpose g2p converters. In this section we briefly explore what further improvements might be possible with a more advanced pronunciation modeling approach.

Imagine that one could automatically convert the baseline transcription of a name into a set of transcriptions always containing the 'true' transcription of that name. How good would the recognition be then? To that end we tested a lexicon containing the four g2p transcriptions of experiment 3, plus all the nativized auditorily verified transcriptions found in the ASNC (this yielded 11.3 transcriptions per name on average). The observed improvements (see Table 7) are substantial for all speaker tongue and name source combinations. It does not come as a surprise that the largest gains are obtained for NN2 speakers reading NN2 names. In spite of the utopic transcription generator that was used, the experiment does indicate that phonological transformation rules which aim to convert g2p transcriptions into 'true' transcriptions may produce a better lexicon. First attempts in this direction already yielded promising results which are described in another paper presented at this conference [11].

Table 7: *NER (%) obtained with AC-MULTI and a lexicon with Dutch, English, French and German g2p transcriptions + all auditorily verified transcriptions found in the ASNC. In bold are gains > 20% w.r.t. the system of experiment 3.*

	Name source			
Speaker origin	**NAT**	**NN1**	**NN2**	**All**
NAT	**3.7**	**2.6**	**1.7**	**3.2**
NN1	**5.3**	**3.7**	**1.9**	**4.6**
NN2	**10.5**	**6.3**	**4.8**	**8.8**
All	**5.8**	**3.7**	**2.6**	**4.9**

5. Conclusions and future work

We have carefully analyzed the impact of pronunciation modeling and acoustic modeling approaches on the performance of automatic name recognition as a function of two dependent variables: the mother tongue of the speaker (the speaker tongue) and the language of origin of the name (the name source). Both variables were investigated according to three language categories: (1) native, (2) non-native but familiar to many native speakers, and (3) non-native and not familiar to native speakers.

Just adding transcriptions of foreign g2p converters and mapping the foreign phonemes to native phonemes leads to substantial gains for names whose source is covered by one of the added g2p converters. The gains comply with the ability of the speakers to use their knowledge of the name source language when reading a name.

Introducing acoustic models trained on multilingual speech data showed additional gains for all non-native name categories, but at the expense of a substantial loss in the recognition of native names uttered by native speakers. We found evidence that most of the improvement actually stems from the fact that the acoustic models are less specialized in representing the native sounds, and therefore beneficial to the recognition of any kind of non-native sounds. Moreover, if the name source is known, the loss of accuracy in the case of native speakers reading native names can be eliminated by excluding foreign transcriptions for these names.

Finally, we demonstrated that substantial further improvements may be possible with a more advanced pronunciation modeling approach that can generate transcriptions, some of which are close to the 'true' transcription of the name.

6. Acknowledgements

The presented work was carried out in the context of the Autonomata Too research project, granted under the Dutch-Flemish STEVIN program.

7. References

[1] Yvon, F.; Boula de Mareüil, P.; dAlessandro, C.; Aubergé, V.; Bagein, M.; Bailly, G.; Béchet, F.; Foukia, S.; Goldman, J.; Keller, E; OShaughnessy, D.; Pagel, V.; Sannier, F.; Véronis, J and Zellner, B. (1998). "Objective evaluation of grapheme to phoneme conversion for text-to-speech synthesis in French", in Computer Speech and Language 12, 393-410.

[2] Cremelie, N. and ten Bosch, L. (2001). "Improving the recognition of foreign names and non-native speech by combining multiple grapheme-to-phoneme converters", in Proc. ISCA ITRW on Adaptation Methods for Speech Recognition, 151-154, Sophia Antipolis, France.

[3] Maison, B.; Chen, S. and Cohen, P. (2003). "Pronunciation modeling for names of foreign origin", in Proc. ASRU, 429-434, Virgin Islands, USA.

[4] Goronzy, S.; Rapp, S. and Kompe, R. (2004). "Generating non-native pronunciation variants for lexicon adaptation", in Speech Communication 42, 109-123.

[5] Bartkova, K. and Jouvet, D. (2007). "On using units trained on foreign data for improved multiple accent speech recognition", in Speech Communication 49, 836-846.

[6] Bonaventura, P.; Gallocchio, F.; Mari, J. and Micca, G. (1998). "Speech recognition methods for non-native pronunciation variants", in Proc. ESCA Workshop on Modeling Pronunciation Variation for ASR, 17-22, Rolduc, The Netherlands.

[7] Van Leeuwen, D. and Orr, R. (1999). "Speech recognition of non-native speech using native and non-native acoustic models", in Proc. workshop MIST, Leusden, The Netherlands.

[8] Stemmer, G.; Nöth, E. and Niemann, H. (2001). "Acoustic modeling of foreign words in a German speech recognition system", in Proc. Eurospeech, 2745-2748, Aalborg, Denmark.

[9] Van den Heuvel, H.; Martens, J-P.; D'hoore, B.; D'Haene, C. and Konings, N. (2008) "The Autonomata Spoken Names Corpus", in Proc. LREC, Marrakech.

[10] Stouten, F. and Martens, J.-P. (2007). "Recognition of foreign names spoken by native speakers", in Proc. Interspeech, 2133-2136, Antwerp, Belgium.

[11] Van den Heuvel, H.; Réveil, B. and Martens, J.-P. (2009), "Pronunciation-based ASR for names", presented at this conference.